基于 matlab 的全最小一乘法

张运胜，高雷阜
（辽宁工程技术大学理学院，辽宁 阜新 123000）

摘要：介绍最小二乘法，最小一乘法与全最小一乘法的差异，全最小一乘法的原理，在基于 matlab 的基础上写出了全最小一乘法的算法的程序，举例说明全最小一乘法的一些应用，得到了在一些情形下使用全最小一乘法求回归直线的优势他是也提供了一种新思想。

关键词：最优化；全最小一乘法；matlab 软件；最小二乘法
中图分类号：O171

Matlab based total least absolute deviation

Zhang Yunshang, Gao Leifu
(college of science, Liaoning Technical University, LiaoNing FuXin 123000)

Abstract: In tthe paper introduced least square method least absolute deviation and total least absolute deviation and discussed their differences and total least absolute deviation of theory; used matlab procedure calculate total least absolute deviation, then illustrated the application on total least absolute deviation, in same case , the new method and the advantage of the linear regression was given by total least absolute deviation .

Keywords: optimization; total least absolute deviation; matlab software; least square method

0 引言

设有 n 个样本点，

$$(x_1,y_1),(x_2,y_2)\ldots(x_n,y_n) \qquad (1)$$

若这些点大致分布在一条狭长的带状区域里，则可以在某一标准用直线 $y=ax+b$ 去拟合它们。我们已经知道在最小二乘法的标准有很多优秀的性质，但是很容易受到奇异点的影响，奇异点在实际的应用中是不可避免的，当原始数据存在奇异点时，应用最小二乘法得到的结果就会产生很大的偏差，其可信度就会大大降低，对于后面的预测也会产生不良的影响，针对这个原因，人们提出处了最小一乘法[1][2]和全最小一乘法[3]去解决奇异点带来的不便，近年来最小一乘法和全最小一乘法受到统计界和优化界许多学者的重视。本文用 matlab 数学软件来设计全最小一乘法的算法，用数学实验来比较在各种标准下，各自的残差绝对值的和与样本点到回归直线距离的和，从而说明在一些情形下使用全最小一乘法的优势所在。

1 全最小一乘法的原理和相关结论

设 $y=ax+b$ 为回归直线，$J_1(a,b)=\sum_{i=1}^{n}|y_i-y_i^*|^2=\sum_{i=1}^{n}|y_i-ax_i-b|^2=\sum_{i=1}^{n}|e_i|^2$ 为回归值，残差为 $J_1(a,b)=\sum_{i=1}^{n}|y_i-y_i^*|^2=\sum_{i=1}^{n}|y_i-ax_i-b|^2=\sum_{i=1}^{n}|e_i|^2$，

最小二乘法的目标函数

基金项目：辽宁省教育厅自然科学基金
作者简介：张运胜，（1882-），男，研究生，运筹控制优化。
通信联系人：高雷阜，（1963-），男，教授，博导，最优化. E-mail: zhangysbad@126.com

$$J_1(a,b)=\sum_{i=1}^{n}|y_i-y_i^*|^2=\sum_{i=1}^{n}|y_i-ax_i-b|^2=\sum_{i=1}^{n}|e_i|^2 \quad (2)$$

为最小，当目标函数 $J_1(a,b)$ 最小时，可以根据它的优良性质， $J_1(a,b)$ 对 a,b 求偏导数，令偏导数为 0.算出 a,b ，得到回归直线 $y=ax+b$ 。一般应用 $|y_i-ax_i-b|$ 来刻画点 (x_i,y_i) 到直线 $y=ax+b$ 的远近程度，但绝对值不便于计算，所以用 $(y_i-ax_i-b)^2$ 来代替 $|y_i-ax_i-b|$，这样就导致了样本点中的奇异点与回归方程有的偏差，其平方后偏差就更大，为使偏差的平方和最小，就不得不将方程与这个奇异点靠进，这样导致所求的方程效果不好，即最小二乘法的不稳健性，为此就引入了最小一乘法，最小一乘法的目标函

$$J_2(a,b)=\sum_{i=1}^{n}|y_i-y_i^*|=\sum_{i=1}^{n}|y_i-ax_i-b|=\sum_{i=1}^{n}|e_i| \quad (3)$$

为最小，可以看出最小一乘法是使残差的绝对值最小，用的是 1 次，不存在缩放的影响，最小一乘法对奇异点的影响较小，但是，最小最小一乘法的计算要用迭代法**[4]**或者线性规划的方法[2][5]， 计算复杂，在最小一乘法的基础上做了改进，几年来人们提出了全最小一乘法，全最小一乘法的目标函数是在最小一乘法的目标函数基础上改进得到的。全最小一乘法的目标函数为

$$\begin{aligned}J_3(a,b)&=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-y_i^*|\\&=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-ax_i-b|=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|e_i|\end{aligned} \quad (4)$$

可以看出它最小的几何意义是刻画点 (x_i,y_i) 到直线 $y=ax+b$ 的距离的和最小，也能很好的处理奇异点的问题，并且比最小一乘法的目标函数的最小值更准确。下面就得讨论它的算法。

定理 1（存在性）[6][7]在全最小一乘法的准则

$$\begin{aligned}J_3(a,b)&=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-y_i^*|\\&=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-ax_i-b|=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|e_i|\end{aligned}$$

下，存在最优的 a,b 或者最优直线 $y=ax+b$

定理 2（必要性）[6][7] 在全最小一乘法的准则

$$\begin{aligned}J_3(a,b)&=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-y_i^*|\\&=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-ax_i-b|=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|e_i|\end{aligned}$$

下，最优直线 $y=ax+b$ 至少经过（1）的 2 个样本点。

可以看出全最小一乘法准则下最优直线不一定是唯一的，定理 2 是一个关键，它为在全最小一乘法准则下，求确切的最优解 $y=ax+b$ 或者 a,b 提供了依据和方法。要求得回归直线 $y=ax+b$ 。我们只要在过 2 样本点的直线中去寻找，直线最多的数目只有

$c_n^2 = n(n-1)/2$ 条，这个计算只有有限步，而且是精确的解，它的算法复杂度为 $o(n^2)$，它的算法比最小一乘法容易的多。

2 全最小二乘法计算步骤及基于 matlab 的程序

由定理 2 可知最优直线至少过 2 个样本点，得到计算程序的步骤：[6]

1 ，求过点 $(x_i, y_i), (x_j, x_j)(1 \le i < j \le n)$ 的直线，计算出直线的系数，

$$a_{ij} = \frac{y_i - y_j}{x_i - x_j} \ (x_i \ne x_j) \ ; b_{ij} = \frac{x_i y_j - x_j y_i}{x_i - x_j} \ (x_i \ne x_j)$$

过 $(x_i, y_i), (x_j, x_j)(1 \le i < j \le n)$ 的直线为

$$\begin{cases} x = x_i & x_j = x_i \\ y = a_{ij} x_i + b_{ij} & x_j \ne x_i \end{cases} \qquad (5)$$

2 计算与 5 式对应的距离和

$$d_{ij} = \begin{cases} \sum_{k=1}^{n} | x_i - x_k | & x_j = x_i \\ \frac{1}{\sqrt{1 + a_{ij}^2}} \sum_{k=1}^{n} | y_k - a_{ij} x_k - b_{ij} | & x_j \ne x_i \end{cases}$$

$$\min d_{ij} \quad \min d_{ij}$$

3,计算 $1 \le i < j \le n$ ，与 $1 \le i < j \le n$ 对应的直线就是最优直线

根据上述步骤得到的用 matlab 软件[8]的全最小一乘法程序如下：

```
x=input('输入横坐标的向量 x:');
y=input('输入纵坐标的向量 y:');
l=input('输入纵坐标的向量 y 最大值分量值:');
n=length(x);
a=zeros(n);
b=zeros(n);
d=l.*ones(n);
for i=1:n
    for j=i+1:n
        a(i,j)=(y(i)-y(j))/(x(i)-x(j));
        b(i,j)=(x(i)*y(j)-x(j)*y(i))/(x(i)-x(j));
        for k=1:n
            s(k)=(1/(1+(a(i,j))^2)^1/2)*(abs(y(k)-a(i,j)*x(k)-b(i,j)));
            s;
        end
        d(i,j)=sum(s);
    end
end
 [c,row]=min(d);
 [c2,col]=min(c);
```

```
row=row(col);
col;
i=row;
j=col;
plot(x,y,'g*',x(i),y(i),'ro',x(j),y(j),'ro')
hold on
m=(y(i)-y(j))/(x(i)-x(j))
n=(x(i)*y(j)-x(j)*y(i))/(x(i)-x(j))
y=m*x+n
plot(x,y,'b')
```

该算法可以直接求出最优直线 $y=ax+b$，画出回归直线经过的 2 个点。

3 基于 matlab 的实例

在统计的样本数据中，可能出现各种异常点，异常点会对回归直线产生影响，从中国统计数据网发布得第一年 6 月到第二年 6 月消费者信心指数数据如表 1 所示

表 1 头年 6 月到下年 6 月消费者信心指数

Tab.1 consumer confidence index from june to next year on june

序号	*1*	2	3	4	5	6	7	8	9	10	11	12	13
月份	6	7	8	9	10	11	12	1	2	3	4	5	6
指数	76	83	93	99	105	114	120	159	122	139	118	126	149

观察一下表中所列数据。容易看出第二年 1 月的消费者信心指数是 200 是一个异常值，可以用 12 月和 2 月的平均值（120+122）/2=121 去修正它并将它视为正常值。

1）用修正后的结果去用最小二乘法和全最小一乘法拟合，得到的回归直线和图形如图 1,2 所示

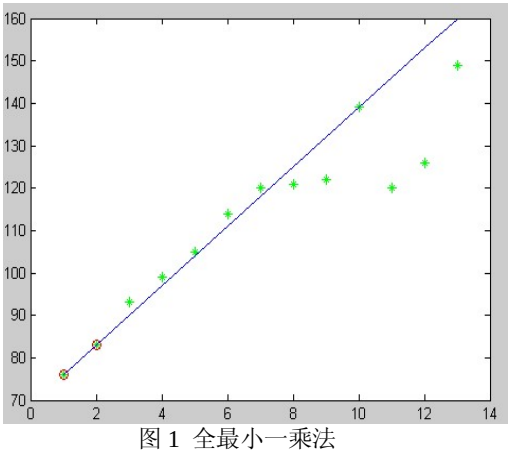

图 1 全最小一乘法

Fig.1 total least absolute deviation

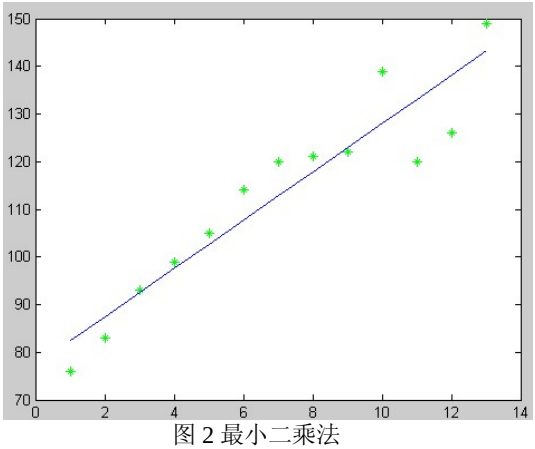

图 2 最小二乘法

Fig.1 least square method

全最小一乘法的回归直线是 $y=7x+69$

最小二乘法的回归直线是 $y=6x+77.5$

2）在 matlab 软件上计算残差绝对值的和与样本点到相应回归直线的距离和比较这 2 种准则的最优性。计算残差绝对值的和的公式[6]为:

$$\sum_{i=1}^{n}|y_i-y_i^*|=\sum_{i=1}^{n}|y_i-ax_i-b|=\sum_{i=1}^{n}|e_i|$$

样本点到相应回归直线的$y=ax+b$的距离和为：

$$(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-y_i^*|$$

$$=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|y_i-ax_i-b|=(1+a^2)^{-\frac{1}{2}}\sum_{i=1}^{n}|e_i|$$

结果计算得到表 2

表 2 残差绝对值的和与样本点到相应回归直线的距离和

Tab.2 The absolute values and the sample points to the regression line distance of the sum

方法与回归直线指标	*最小二乘法* $y=6x+77.5$	*全最小一乘法* $y=7x+69$
残差绝对值的和	*25.20*	*22.2500*
距离和	*1.400*	*0.8900*

通过图形可以看出全最小二乘法经过第 1，第 2 个点过这 2 个点的直线是最优直线，通过表 2 可以知道用全二次最小一乘法可以得到比最小二乘法好的最优直线。

4 结论

本文给出了最小二乘法与全最小一乘法基于 matlab 的数学实验图形。最小二乘法有很好的解析性质，使得最小二乘法称为普遍接受的求回归直线的方法，但最小二乘法受“奇异点”的影响，不能很好的对其进行拟合，本文采用全最小一乘法能对有奇异点的样本进行很好的拟合，却没有最小一乘法的计算复杂，在全最小一乘法的标准下，克服了最小二乘法对的奇异点困难，又没有陷入最小一乘法的计算困难，在 matlab 的软件下很好的能编出全最小一乘法的程序。用数学实验说明使用全最小一乘法求回归直线具有一定的优势，也是我们考虑线性回归直线的一种好的思路。

[参考文献] (References)

[1] 陈希孺, 赵林城. 线性模型中的 M 方法[M]. 上海: 上海科学技术出版社, 1996.
[2] 万树平，基于最小一乘估计的多传感器信息融合方法[J],计算机工程，2010，36（2）：257-259
[3] 吴克法，关于加权全最小一乘法的探讨[J]，应有数学学报，2002 25（3）：439-447
[4] 杨桂元．残差绝对值之和最小准则下回归方程的求法[j] ，运筹与管理 2001，10(1)：20-23.
[5] 苗增强, 基于最小一乘法的组合赋权法在中长期负荷预测中的应用[J],D 电力系统保护与控制，2009,37（2）28-32
[6] 冯守平，关于加权全最小一乘法[J].应用概率统计，2009(2):135-142
[7] 洪文，吴本忠，LING4.0 for windows 最优化软件及其应用[M],北京：北京大学出版社，2001
[8] 张贤明，MATLAB 语言及应用案例[M],南京：东南大学出版社,2010